DOCUMENT RESUME

ED 380 504 TM 022 867

AUTHOR Messick, Samuel
TITLE Alternative Modes of Assessment, Uniform Standards of Validity. Research Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-60
PUB DATE Dec 94
NOTE 26p.; Paper presented at a Conference on Evaluating Alternatives to Traditional Testing for Selection (Bowling Green, OH, October 25-26, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; *Educational Assessment; Inferences; Multiple Choice Tests; *Psychological Testing; Scores; *Standards; *Test Interpretation
IDENTIFIERS *Alternative Assessment; *Performance Based Evaluation

ABSTRACT

In contrast to multiple choice, alternative modes of assessment afford varying degrees of openness in the allowable responses. Prominent among the alternatives is the assessment of performance, sometimes in its own right where the issue is the quality of the particular performance per se, but more often as a vehicle for the assessment of knowledge, skill, or other attributes. Because inferences about score meaning in construct terms and about the action implications of that meaning are fundamentally similar in the alternative assessment modes (despite surface differences), the same standards of validity apply to all educational and psychological measurement. These standards are addressed in terms of content, substantive, structural, generalizability, external, and consequential aspects of construct validity. (Contains 42 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-60

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ALTERNATIVE MODES OF ASSESSMENT, UNIFORM STANDARDS OF VALIDITY

Samuel Messick

ED 380 504

ETS

Educational Testing Service
Princeton, New Jersey
December 1994

TM022867

Copyright © 1994. Educational Testing Service. All rights reserved.

ALTERNATIVE MODES OF ASSESSMENT, UNIFORM STANDARDS OF VALIDITY

Samuel Messick
Educational Testing Service

ALTERNATIVE MODES OF ASSESSMENT, UNIFORM STANDARDS OF VALIDITY

Samuel Messick
Educational Testing Service

ABSTRACT

In contrast to multiple choice, alternative modes of assessment afford varying degrees of openness in the allowable responses. Prominent among the alternatives is the assessment of performance, sometimes in its own right where the issue is the quality of the particular performance per se, but more often as a vehicle for the assessment of knowledge, skill, or other attributes. Because inferences about score meaning in construct terms and about the action implications of that meaning are fundamentally similar in the alternative assessment modes (despite surface differences), the same standards of validity apply to all educational and psychological measurement. These standards are addressed in terms of content, substantive, structural, generalizability, external, and consequential aspects of construct validity.

ALTERNATIVE MODES OF ASSESSMENT, UNIFORM STANDARDS OF VALIDITY[1]

Samuel Messick
Educational Testing Service

Nowadays, when people speak of alternatives to traditional testing, they are really referring to alternatives to standardized paper-and-pencil multiple-choice testing. In particular, the critical concern is with alternatives to the multiple-choice format because of possible constraints on the kinds of thinking and higher-order cognitive processing that can be assessed (N. Frederiksen, 1989). The issues of standardization and paper-and-pencil delivery are not as salient. This is so because alternatives to paper-and-pencil presentation, such as by interviewer or computer, have only subtle measurement implications as long as what is presented are still multiple-choice items. The measurement implications of the delivery mode are important but are highly dependent on the assessment mode. Nor is it the case that the alternative modes of assessment should be unstandardized, although the conditions that are controlled may differ from those of traditional testing (Messick, 1993). Hence it is alternatives to multiple choice that are our main concern here.

The defining feature of a multiple-choice item is that the respondent must select an answer from among a set of options. The obvious alternative requires the respondent to construct the answer de novo. This constructed-response alternative has led to a resurgence of interest in performance assessment, which, though long a staple of industrial and military applications, is becoming increasingly popular in educational settings, especially in connection with standards-based education reform.

---

[1] This paper was presented at a Conference on Evaluating Alternatives to Traditional Testing for Selection sponsored by Bowling Green State University, October 25-26, 1994. Acknowledgements are gratefully extended to Randy Bennett. Ann Jungeblut, Donald Powers, and William Ward for their reviews of the manuscript.

Because performance assessments appear to be noticeably different from traditional testing, the question arises as to whether the same standards of validity should apply to them as opposed to specialized validity standards (Linn, Baker, & Dunbar, 1991; Messick, 1994; Moss, 1992). Given the developing consensus that validity is a unified concept (APA, 1985; Messick, 1989), the preference would appear to favor uniform validity standards for all educational and psychological assessments, including performance assessments. Indeed, this paper attempts to justify this position in terms of a comprehensive view of construct validity. However, first we must examine some varied perspectives on the meaning of performance assessment, because different conceptions have distinctly different implications for validation.

ASSESSMENT OF PERFORMANCE OR OF CONSTRUCTS?

First we consider some variable properties of performance assessments in contradistinction to multiple choice. Next, we examine the tension between task-driven and construct-driven performance assessment in terms of whether the performance is to serve as the target or the vehicle of the assessment. Then we highlight the two major sources of test invalidity because they provide a basis for determining what "authenticity" and "directness" of performance assessment might mean in validity terms.

Conceptions of Performance Assessment

In essence, a performance assessment requires the respondent to execute a task or process and bring it to completion (Wiggins, 1993). That is, the examinee performs, creates, or produces something over a sufficient duration of time to permit evaluation of either the process or the product, or both. This is in contradistinction to the impoverished trace or scorable record resulting when one merely marks a correct or preferred option on an answer sheet as in a multiple-choice test, which does not reflect the amount or kind of thinking or effort that may underlie the choice of option.

Indeed, with respect to task processing, the boundary between multiple-choice tests and performance assessments is a fuzzy one because some respondents on many multiple-choice items and most respondents on difficult multiple-choice items execute the solution process as a means of selecting the

appropriate option (Traub, 1993). A more critical distinction is that the selected option can only be appraised for correctness or goodness with respect to a single criterion. There is no record, as in the typical performance assessment, of an extended process or product that can be scored for multiple aspects of quality.

A further complication is that the contrast between multiple-choice items and open-ended performance tasks is not a dichotomy, but a continuum representing different degrees of structure versus openness in the allowable responses. This continuum is variously described as ranging from multiple-choice to examinee-constructed products or presentations (Bennett, 1993), for example, or from multiple-choice to demonstrations and portfolios (Snow, 1993). Successive intervening stages include items requiring reordering or rearranging, substitution or correction, simple completion or cloze procedures, short essays or complex completions, problem exercises or proofs, teach-back procedures, and long essays.

There is a wide array of structured item formats toward the multiple-choice end of the continuum. For example, Wesman (1971) describes three varieties of the short-answer form, five varieties of the alternate-choice form, two of the matching form, and eight of multiple-choice, including those allowing more than one right answer. In addition, he discusses three types of context-dependent item sets (the pictorial form, the interlinear form, and the interpretive exercise), to which a fourth type (the problem-solving scenario) has been added (Haladyna, 1992). Thus, contingent sets of structured items can be developed to tap complex aspects of task functioning, such as problem-solving processes and strategies (Ebel, 1984) as well as stylistic learning preferences (Heath, 1964). It should be noted that, contrary to popular misconceptions, structured item formats are not limited to the measurement of fact retrieval. They are also used effectively to assess knowledge application, evaluation skills, and problem-solving proficiencies. Multiple- or forced-choice techniques have also been applied in the measurement of social attitudes, personal needs and motives, vocational interests, aesthetic preferences, and human values (Messick, 1979).

In addition, there are a number of formats at intermediate levels of the continuum, one example being multiple-choice items that require the respondent to give reasons why the chosen option is correct and possibly why each of the

unchosen options is incorrect. Another instance is a multiple-rating format in which each of several options is judged for quality against complex standards (Scriven, 1994). Specifically, the respondent might be asked to read a passage for main idea and then to rate each of four sentences -- say, by marking boxes labeled A to F -- for the quality and completeness with which each captures the main idea. An added requirement might be that if none of the statements receives a grade of B or better, the respondent should write an A-quality main idea sentence of his or her own.

It should be noted that this continuum refers to *response*-form, representing various degrees of structure or constraint imposed on the examinee's responses. There is another, at least partly independent, continuum referring to *stimulus*-form that represents various degrees of structure in the questions or problems presented. These two continua are clearly separable in the structured-stimulus direction because highly structured problems can be presented in either multiple-choice or open-ended formats. The question is the degree to which the two continua are also separable in the unstructured-stimulus direction. In this regard, we should explore the possibility of retaining the efficiency of structured or partly structured responses while simultaneously relaxing the degree of structure in the problems posed. As an instance, patient-management problems might be presented with multiple-choice or key-list options at each decision point. The intent would be to create more realistic, less well-structured problems -- perhaps even ill-structured problems -- having structured or semi-structured response formats.

Apart from multiple-choice, the remainder of the response continuum is referred to as involving "constructed responses." However, not all constructed responses -- notably those involving rearranging, substitution, and simple completion -- are properly considered to be performance assessments because they do not yield a scorable record of an extended process or product.

Prototypical performance assessments occur more toward the unstructured end of the response continuum and include such exemplars as portfolios of products collected over time, exhibits or displays of knowledge and skill, open-ended tasks with no single correct approach or answer, hands-on experimentation, and work samples. The openness with respect to response possibilities enables examinees to exhibit skills that are difficult to tap

within the predefined structures of multiple-choice, such as shaping or restructuring a problem, defining and operationalizing variables, manipulating conditions, and developing alternative problem approaches.

Evaluations of performance on such open-ended tasks usually rely on the professional judgment of the assessor, and some proponents view such subjectivity of scoring to be the hallmark of performance assessment (e.g., J. R. Frederiksen & Collins, 1989; Stiggins, 1991). However, this view appears too restrictive because some performance tasks can be objectively scored and some scoring judgments are amenable to expert-system computer algorithms (e.g., Bejar, 1991; Sebrechts, Bennett, & Rock, 1991).

A more likely hallmark of educational performance assessments is their nearly universal focus on higher-order thinking and problem-solving skills. According to Baker, O'Neil, and Linn (1993), "virtually all proponents of performance-based assessment intend it to measure aspects of higher-order thinking processes" (p. 1211). Indeed, performance assessments in education frequently attempt to tap the complex structuring of multiple skills and knowledge, including basic as well as higher-order skills, embedded in realistic or otherwise rich problem contexts that require extended or demanding forms of reasoning and judgment. In this regard, Wiggins (1993) views "authentic" performance assessments as tapping understanding or the application of good judgment in adapting knowledge to fashion performances effectively and creatively.

This mention of "authentic" assessments broaches a further distinction. Just as performance assessments are a more open-ended subset of constructed responses, so-called authentic assessments are a more realistic subset of performance assessments. In particular, authentic assessments pose engaging and worthy problems (usually involving multistage tasks) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world. The assessment challenge of complex performance tasks in general and authentic tasks in particular revolves around issues of scoring, interpretation, and generalizable import of key aspects of the complex performance, especially if the task is not completed successfully.

In performance assessment, one might start by clarifying the nature of the higher-order competencies or other constructs to be assessed and then

select or construct tasks that would optimally reveal them. Or, contrariwise, one might start with an important task that is worthy of mastery in its own right and ask what competencies or other constructs this task reveals. This contrast embodies a tension in performance assessment between construct-centered and task-centered approaches (Messick, 1994). However, what is critical in performance assessment is not what is operative in the task performance but what is captured in the test score and interpretation. Hence, the validity of the construct interpretation needs to be addressed sooner or later in either approach, as does the nature of convergent and discriminant evidence needed to sustain that validity.

Construct-Driven Versus Task-Driven Performance Assessment

The task-centered approach to performance assessment begins by identifying a worthy task and then determining what constructs can be scored and how. Often the mastery of such a worthy task functions as the target of the assessment in its own right, as opposed to serving as a vehicle for the assessment of knowledge, skills, or other constructs. This might occur, for example, in an arts contest or an Olympic figure-skating competition or a science fair. In such cases, replicability and generalizability are not at issue. All that counts is the quality of the performance or product submitted for evaluation, and the validation focus is on the judgment of quality. But note that in this usage of performance assessment as target, inferences are not to be made about the competencies or other attributes of the performers, that is, inferences from observed behavior to constructs such as knowledge and skill underlying that behavior. The latter type of inference requires convergent and discriminant evidence for support.

Large-scale educational projects such as dissertations are often treated as targets in this manner, by crediting the complex accomplishment as meeting established standards with no requirement of predictiveness or domain generalizability (Baker et al., 1993). However, action implications of such complex assessments usually presume, with little or no specific evidence, that there is a global prediction of future success, that the knowledge and skills exhibited in the assessment will enable the student to accomplish a range of similar or related tasks in broader settings.

In contrast, such presumptions should be buttressed by empirical evidence in the performance assessment of competencies or other constructs -- that is, whenever the performance is the vehicle not the target of assessment. A major form of this evidence bears on generalizability and transfer which, as we shall see, represent critical aspects of construct validity. In effect, the meaning of the construct is tied to the range of tasks and situations that it generalizes and transfers to.

The task-centered approach to performance assessment is in danger of tailoring scoring criteria and rubrics to properties of the task and of representing any educed constructs in task-dependent ways that might limit generalizability. In contrast, the nature of the constructs in the construct-centered approach guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. Focussing on constructs also alerts one to the possibility of construct-irrelevant variance that might distort either the task performance or its scoring, or both (Messick, 1994). The task-centered approach is not completely devoid of constructs, of course, because task selection is often influenced by implicit construct notions or informal theories of learning and performance. The key issue is the extent to which the constructs guide scoring and interpretation and are explicitly linked to evidence supporting that interpretation as well as discounting plausible rival interpretations.

Sources of Invalidity

Construct-irrelevant variance is one of the two major threats to validity, the other being construct underrepresentation. A fundamental feature of construct validity is ***construct representation***, whereby one attempts to identify through cognitive-process analysis or research on personality and motivation the theoretical mechanisms underlying task performance, primarily by decomposing the task into requisite component processes and assembling them into a functional model or process theory (Embretson, 1983; Wiley, 1991). Relying heavily on the cognitive psychology of information processing, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge (including metacognitive or self-knowledge) that are implicated in task performance.

In the threat to validity known as "construct underrepresentation," the assessment is too narrow and fails to include important dimensions or facets of the construct. In the threat to validity known as "construct-irrelevant variance," the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct. Both threats are operative in all assessment. Hence a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance.

The concept of construct-irrelevant variance is important in all educational and psychological measurement, including performance assessments. This is especially true of richly contextualized assessments and authentic simulations of real-world tasks. This is the case because, "paradoxically, the complexity of context is made manageable by contextual clues" (Wiggins, 1993, p. 208). And it matters whether the contextual clues that are responded to are construct-relevant or represent construct-irrelevant difficulty.

However, what constitutes construct-irrelevant variance is a tricky and contentious issue (Messick, 1994). This is especially true of performance assessments, which typically invoke constructs that are higher-order and complex in the sense of subsuming or organizing multiple processes. For example, skill in communicating mathematical ideas might well be considered irrelevant variance in the assessment of mathematical knowledge (although not necessarily vice versa). But both communication skill and mathematical knowledge are considered relevant parts of the higher-order construct of mathematical power according to the content standards developed by the National Council of Teachers of Mathematics. It all depends on how compelling the evidence and arguments are that the particular source of variance is a relevant part of the focal construct as opposed to affording a plausible rival hypothesis to account for the observed performance regularities and relationships with other variables.

Authenticity and Directness As Validity Standards

Two terms that appear frequently, and usually in tandem, in the literature of performance assessment are "authentic" and "direct" assessment. They are most often used in connection with assessments involving realistic simulations or criterion samples. If authenticity and directness are

important to consider when evaluating the implications of assessment, they constitute tacit validity standards, so we need to address what the labels "authentic" and "direct" might mean in validity terms.

The major measurement concern of authenticity is that nothing important has been left out of the assessment of the focal construct (Messick, 1994). This is tantamount to the familiar validity standard of minimal construct underrepresentation. However, although authenticity implies minimal construct underrepresentation, the obverse does not hold. This is the case because minimal construct underrepresentation does not necessarily imply the close simulation of real-world problems and resources typically associated with authenticity in the current educational literature on performance assessment. In any event, convergent and discriminant evidence is needed to appraise the extent to which the ostensibly authentic tasks represent (or underrepresent) the constructs they are interpreted to assess.

The major measurement concern of directness is that nothing irrelevant has been added that distorts or interferes with construct assessment. This is tantamount to the familiar validity standard of minimal construct-irrelevant variance (Messick, 1994). Incidentally, the term "direct assessment" is a misnomer because it always promises too much. In education and psychology, "all measurements are indirect in one sense or another" (Guilford, 1936, p. 3). Measurement always involves, even if only tacitly, intervening processes of judgment, comparison, or inference.

UNIFORM VALIDITY STANDARDS

Although on the surface there appear to be a number of differences between performance assessments and traditional multiple-choice testing, the inferences drawn from such alternative modes of assessment, as well as their action implications, are fundamentally similar. Indeed, "there is no absolute distinction between performance tests and other classes of tests" (Fitzpatrick & Morrison, 1971, p. 238). This implies that the same standards of validity should be applied to performance assessments as to all educational and psychological assessments. This is so because what is to be validated is not the test or observation device as such but rather the inferences derived from test scores or other indicators (Cronbach, 1971) -- inferences about score

meaning or interpretation and about the implications for action that the interpretation entails. In essence, then, test validation is empirical evaluation of the meaning and consequences of measurement.

Perennial Validity Questions

To evaluate the meaning and consequences of measurement is no small order, however, and requires attention to a number of persistent validity questions, such as:

- Are we looking at the right things in the right balance?
- Has anything important been left out?
- Does our way of looking introduce sources of invalidity or irrelevant variance that bias the scores or judgments?
- Does our way of scoring reflect the manner in which domain processes combine to produce effects and is our score structure consistent with the structure of the domain about which inferences are to be drawn or predictions made?
- What evidence is there that our scores mean what we interpret them to mean, in particular, as reflections of personal attributes having plausible implications for educational, personnel, or therapeutic action?
- Are there plausible rival interpretations of score meaning or alternative implications for action and, if so, by what evidence and arguments are they discounted?
- Are the judgments or scores reliable and are their properties and relationships generalizable across the contents and contexts of use as well as across pertinent population groups?
- Are the value implications of score interpretations empirically grounded, especially if pejorative in tone, and are they commensurate with the score's trait implications?
- Do the scores have utility for the proposed purposes in the applied settings?
- Are the scores applied fairly for these purposes?
- Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side-effects?

Which, if any, of these questions is unnecessary to address in justifying score interpretation and use? Which, if any, can be forgone in validating the interpretation and use of performance assessments or other alternative modes of assessment? The general thrust of such questions is to seek evidence and arguments to discount the two major threats to construct validity -- namely, construct underrepresentation and construct-irrelevant variance -- as well as to evaluate the action implications of score meaning.

Aspects of Construct Validity

Such questions are inherent in the notion of validity as a unified concept. Unified validity does not imply answering only one overarching validity question or even several questions separately or one at a time. Rather, it implies an integration of multiple supplementary forms of convergent and discriminant evidence to answer an interdependent set of questions. To make this explicit, it is illuminating to differentiate unified validity into several distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied use.

In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989). They are briefly characterized as follows:

- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989).
- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
- The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957).

- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test-criterion relationships (Hunter, Schmidt, & Jackson, 1982).

- The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965).

- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989).

A key issue for the content aspect of construct validity is the specification of the boundaries of the construct domain to be assessed -- that is, determining the knowledge, skills, and other attributes to be revealed by the assessment tasks. The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and especially domain theory, that is, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed.

However, it is not sufficient merely to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to insure that all important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity.

The substantive aspect of construct validity emphasizes two important points: One is the need for tasks providing appropriate sampling of domain

processes in addition to traditional coverage of domain content; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance. Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Embretson, 1983; Loevinger, 1957; Messick, 1989).

According to the structural aspect of construct validity, scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). That is, the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria and rubrics. Ideally, the manner in which behavioral instances are combined to produce a score should rest on knowledge of how the processes underlying those behaviors combine dynamically to produce effects. Thus, the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989).

The concern that a performance assessment should provide representative coverage of the content and processes of the construct domain is meant to insure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

The emphasis here is on generalizability in two senses, namely, as it bears on reliability and on transfer. Generalizability as reliability refers to the consistency of performance across the raters, occasions, and tasks of a particular assessment, which might be quite limited in scope. For example, we have all been concerned that some assessments with a narrow set of tasks might attain higher reliability in the form of cross-task consistency, but at the expense of construct validity. In contrast, generalizability as transfer

requires consistency of performance across tasks that are representative of the broader construct domain. That is, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956). Thus, generalizability as transfer depends not only on generalizability theory but also on construct theory. In essence, then, generalizability evidence is an aspect of construct validity because it establishes boundaries on the meaning of the construct scores.

However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. This conflict between depth and breadth of coverage is often viewed as entailing a trade-off between validity and reliability (or generalizability). It might better be depicted as a trade-off between the valid description of the specifics of a complex task performance and the power of construct interpretation. In any event, such a conflict signals a design problem that needs to be carefully negotiated in performance assessment (Wiggins, 1993).

The external aspect of construct validity refers to the extent to which the assessment scores' relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations.

Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.

The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score

interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation or with unfairness in test use. However, this form of evidence should not be viewed in isolation as a separate type of validity, say, of "consequential validity." Rather, because the values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1980).

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989). That is, low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected persons to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence. In contrast, if adverse consequences are associated with valid measurement, the primary concern is one of social policy that weighs those adverse consequences against potential benefits in deciding whether to use the test or alternative modes of assessment.

From the discussion thus far, it should be clear that test validity cannot ***rely*** on any one of the supplementary forms of evidence just discussed. However, neither does validity ***require*** any one form, granted that there is defensible convergent and discriminant evidence supporting score meaning. To the extent that some form of evidence cannot be developed -- as when criterion-related studies must be forgone because of small sample sizes, unreliable or contaminated criteria, and highly restricted score ranges -- heightened emphasis can be placed on other evidence, especially on the construct validity of the predictor tests and the relevance of the construct to the criterion domain (Guion, 1976; Messick, 1989). What *is* required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence had to be forgone. Hence, validity becomes a unified concept and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity.

Validity As Integrative Summary

The six aspects of construct validity apply to all educational and psychological measurement, including performance assessments or other alternative assessment modes. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. In previous writings I maintained that it is "the relation between the evidence and the inferences drawn that should determine the validation focus" (Messick, 1989. p. 16). This relation is embodied in theoretical rationales or persuasive arguments that the obtained evidence both supports the preferred inferences and undercuts plausible rival inferences. From this perspective, as Cronbach (1988) concluded, validation is evaluation argument. That is, as stipulated earlier, validation is empirical evaluation of the meaning and consequences of measurement. The term "empirical evaluation" is meant to convey that the validation process is scientific as well as rhetorical and requires both evidence and argument.

By focussing on the argument or rationale employed to support the assumptions and inferences invoked in the score-based interpretations and actions of a particular test use, one can prioritize the forms of validity evidence needed in terms of the important points in the argument that require justification or support (Kane, 1992; Shepard, 1993). Helpful as this may be, there still remain problems in setting priorities for needed evidence because the argument may be incomplete or off target, not all the assumptions may be addressed, and the need to discount alternative arguments evokes multiple priorities. This is one reason that Cronbach (1989) stressed cross-argument criteria for assigning priority to a line of inquiry, such as the degree of prior uncertainty, information yield, cost, and leverage in achieving consensus.

The point here is that the six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases and, if not, requiring that an argument be provided that such omissions are defensible. They are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that

attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly.

That is, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and are generalizable across a range of tasks, settings, and occasions. Furthermore, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of presumed or documented relationships with nontest behaviors and anticipated outcomes or consequences. The challenge in test validation is to link these inferences to convergent evidence supporting them as well as to discriminant evidence discounting plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else provide compelling reasons why not, which is what is meant by validity as a unified concept -- a concept that applies with equal force not only to traditional tests but also to alternative modes of assessment.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210-1218.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522-532.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and public policy -- Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147-171). Chicago: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Ebel, R. L. (1984). Achievement test items: Current issues. In B. S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage*. (pp. 141-154). Hillsdale, NJ: Erlbaum.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology*, *10*, 121-131.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 237-270). Washington, DC: American Council on Education.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27-32.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193-202.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 777-828). Chicago: Rand McNally.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice*, 11(1), 21-25.

Heath, R. W. (1964). Curriculum, cognition, and educational development. *Educational and Psychological Measurement*, 24, 239-253.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies*. San Francisco: Sage.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, *16*, 294-304.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 15-21.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694 (Monograph Supplement 9).

Messick, S. (1979). Potential uses of noncognitive measurement in education. *Journal of Educational Psychology*, 71, 281-292.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1993). Trait equivalence as construct validity across multiple methods of measurement. In R. E. Bennett & W. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-73). Hillsdale, NJ: Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243-299). Hinsdale, IL: Dryden Press.

Scriven, M. (1994). Death of paradigm: Replacing multiple choice with multiple ratings. Paper presented to the Annual Meeting of the American Educational Research Association, New Orleans, April 4-8.

Sebrechts, M. M., Bennett, R. E., & Rock D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856-862.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education, 19*, 405-450.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-396.

Snow, R. E. (1993). Construct validity and constructed response tests. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Erlbaum.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education, 4*, 263-273.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward, Jr. (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Erlbaum.

Wesman, A. G. (1971). Writing the test item. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 81-129). Washington, DC: American Council on Education.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 83*, 200-214.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.